Re-Architecting DRAM Memory Systems with 3D Integration and Photonic Interfaces

Ishan G Thakkar, Sudeep Pasricha
Department of Electrical and Computer Engineering
Colorado State University, Fort Collins, CO, U.S.A.
{ishan.thakkar, sudeep}@colostate.edu

Abstract - In recent years, due to well-known "memory-wall" limitations, DRAM latency and energy/bit characteristics have not improved as rapidly as DRAM capacity and bandwidth with technology scaling. Despite plenty of research efforts during the last decade to overcome the hurdles in DRAM scaling, high performance computing systems of the future will still require improvements in DRAM latency and bandwidth within stringent power constraints. In this paper, firstly we establish the need for reinventing DRAM architectures by identifying the limiting features of several state-of-the-art DRAM architectures. Then we discuss some useful features of 3D integration and photonic interfaces that have the potential to greatly improve the performance and energy of future DRAMs. Finally, we present a novel optically-interfaced 3D-DRAM architecture with 3D data organization, which achieves 4.4× improvement in performance and 83.8% reduction in per-bit energy on average over 3D-DRAM architectures from prior work.

I. INTRODUCTION

In recent years, DRAM latency and bandwidth have not improved as rapidly as DRAM capacity with shrinking process technology, owing to well-known "memory-wall" limitations [1]. The traditional means of improving memory access performance by increasing clock frequency is also no longer practical due to increasingly stringent power constraints that limit further frequency scaling. Moreover, the latency-tolerance techniques prevalent today, such as multi-level caches, row prefetching, burst mode access, and memory parallelism [2], [3], are not expected to scale well in terms of latency and power for high performance computing systems of the future [4]. These trends are forcing designers to reinvent DRAM architectures, to overcome the hurdles in DRAM performance and power scaling. Consequently, today, many researchers invest their efforts in forging new methods for DRAM cell organization [5], [28], memory request scheduling [6], [7] and DRAM refresh [8], [9]. However, there still remains a critical unmet need for an innovative DRAM architecture that can provide simultaneous improvements in latency, bandwidth, power and cost.

Since the emergence of TSV (through silicon via) based 3D integration technology, 3D-stacked DRAM has been a promising option to alleviate many of the limitations of commodity 2D-DRAMs. The main advantage of TSV-based stacking is that it reduces the wire-length between modules located on different tiers, which in turn reduces delay and energy of inter-module interconnects. In recent years, a few 3D DRAM architectures have been proposed [10]-[12], [28] that have exploited the benefits of TSVs with fine-grained partitioning and activation to notably improve memory performance and parallelism. Such 3D-stacked DRAMs require new methods for efficient address and data path routing and 3D cell organization, to realize their potential and achieve improved memory parallelism and latency. But, TSV-based 3D systems suffer from low yield and thermal/noise issues which affect the productization of these systems. Therefore, the commercial success and performance benefits of TSV-based 3D integration for future DRAM systems depend on the density and interconnect level of TSVs employed.

In this paper, we identify the fundamental elements of TSV-based 3D integration technology that would have very significant influence on how future 3D DRAM systems will evolve. Furthermore, due to pin-bandwidth limitations of the traditional DRAM interface, the internal memory bus and I/O bus are shared among all the banks in state-of-the-art DRAM modules. This causes bus contention outside of banks, limiting the benefits of increased bank-level concurrency in these modules. Therefore, increasing concurrency within a memory system has limited benefits unless the pin-bandwidth limitation of DRAM interface is alleviated. We therefore also discuss how an intelligent layout of TSV buses at the bank-level granularity and a high speed photonic interface can be used to alleviate the pin-bandwidth limitations for future DRAM interfaces. We introduce a novel optically-interfaced 3D-DRAM architecture with new methods of 3D data organization and TSV layout, which can become a promising solution for high-bandwidth, low-latency, and low-energy DRAMs of the future.

II. BACKGROUND, TRENDS AND OPPORTUNITIES

In this section, we identify the fundamental elements of the state-of-the-art 2D-DRAM and 3D-DRAM architectures that contribute significantly to the overall performance, energy consumption, and die area. Then we briefly discuss the trends and opportunities of the TSV-based 3D integration technology and photonic interface technology in order to determine the roadmap of future 3D-DRAM architectures.

The fundamental factor simultaneously affecting energy and latency of DRAM systems is RC loading of the memory access path. The memory access path has three stages in the hierarchy, with 2D routing paths/lines at the top, global lines in the middle, and local wordlines and bitlines at the bottom. The RC loading of 2D routing lines and global lines is proportional to their length. In contrast, the RC loading of the local wordlines and bitlines is proportional to their respective count. The research efforts reported in [13] and [14] show that reducing the number of bitlines and wordlines per subarray significantly reduce the RC loading, yielding greater performance with lower energy consumption. But, such a reduction can significantly increase the area overhead, harming overall area efficiency, if it is done without due deliberation. Consequently, some modern DRAM designs reported in [14] and [15] have optimally reduced the number of bitlines and wordlines in a subarray to be 256 so that it does not significantly increase the area overhead.

Unlike 2D-DRAMs, the designers of 3D-stacked DRAMs can use the extra (vertical) dimension to reduce RC loading of 2D routing paths and global lines. In 3D-stacked DRAM, the address and data buses are routed to each die layer in the vertical direction using TSVs, and after reaching a die they are routed to the edges of individual banks along 2D routing paths. These 2D routing paths are usually realized in intermediate metal layers, which contributes significantly to the overall latency, energy, and area costs. From the edges of individual banks, the pre-decoded address lines are routed to individual

subarrays along the global word lines and global column select lines. At a subarray level, the global address lines are further decoded to drive local word lines and local column select lines. Similarly, data buses reach individual subarrays as local data lines, which are driven by global data lines. In these global data lines, the repeaters and drivers incur extra overhead of latency and energy consumption at each die. Modeling these structures using the CACTI-3DD [23] tool indicates that a 2Gb DRAM die at the 45nm technology node has approximate dimensions of 8mm × 12mm. For this die, an average length of the 2D routing path would be $\sqrt{8 \times 12}$ = 9.8mm (geometric mean of die), which translates into 82.3pJ energy and 0.8ns delay for one 2D routing wire (M3 layer). In contrast, if this 2D routing is replaced by vertical routing using 4-tier long TSVs, then the energy and delay of one routing wire becomes 0.64pJ and 0.12ns respectively. This analysis indicates a substantial reduction in delay and energy for TSV-based vertical routing compared to 2D planar routing. Thus integrating TSV-based routing in 3D-DRAM architectures can provide immense performance and energy consumption benefits. We now summarize a few recent works that exploit this idea.

Chen et al. [23] discussed the pros and cons of coarse-grained and fine-grained rank-level partitioning for 3D DRAMs with respect to latency, energy consumption, and area efficiency. For the coarse-grained rank-level cell organization, the inter-bank communication within one rank occurs via 2D routing paths. Kang et al. [11] proposed extending the commodity DDR3 architecture to 3D-DDR3 by employing the coarse-grained rank-level cell organization and showed benefits in performance due to the reduced length and latency of a TSV-based memory bus. Micron's hybrid memory cube (HMC) exploits fine-grained rank-level partitioning to further reduce the access latency and increase memory parallelism [12]. The analysis in [23] indicated that the activation energy and latency of the fine-grained design are reduced by 48.5% and 46.9% respectively compared to the coarse-grained design, because of the reduced bank size and optimized data path routing in the fine-grained design. The total die area for the fine-grained design reduces by 35.9% in spite of a 3.7% TSV area overhead. Also, the latency of the internal memory bus is reduced by 62.8% for the fine-grained design. The reason behind the improved results for the fine-grained model is that the model utilizes the potential bandwidth of TSVs for inter-bank transfers, which eliminates the need for 2D routing paths for such transfers and alleviates related overheads. It is worth noting that HMC's fine-grained rank-level partitioning uses global lines to route address and data bits to individual sub-arrays of a bank. Intuitively, one can conjecture that by intelligently using TSVs to route data and address lines to individual subarrays, the need to use global lines at each bank can be eliminated, which would further minimize overall latency and energy.

Based on the analysis done in recent work, it is obvious that TSV-based aggressive vertical routing of data and address lines has great potential to overcome the hurdles in performance scaling of future DRAMs. But considering the low-yield, high cost and grave thermal/noise issues associated with TSV-based 3D-stacking technology, many DRAM designers have questioned the commercial feasibility of future 3D DRAMs employing TSVs routed directly to individual subarrays. There are a few emerging solutions that show promise to overcome challenges traditionally associated with 3D-stacking. One such solution is Tezzaron's Tungsten-based 1 micron wide TSVs called SuperContacts [16]. SuperContacts can be arrayed on a 2 micron or smaller pitch which translates into 100× greater density for SuperContacts compared to conventional Cu-based TSVs. Such a high density of TSVs enables finer-grained repair at the system-level, increasing the net yield for 3D-IC fabrication. Moreover, SuperContacts are very nearly perfectly thermally compatible with Silicon, which alleviates the effect of thermal-mechanical stress around TSVs, increasing the thermal stability of the system. However, even SuperContacts would incur a high area overhead if they are used to route address and data bits directly to individual subarrays. We propose an interesting solution to this challenge in future 3D DRAM architectures: using SuperContacts coupled with fanout buffers. The details of our proposed architecture are given in Section III.

Another fundamental limiting factor that is being encountered with DRAM scaling is the pin-bandwidth limitation of the conventional electrical interface of commodity DRAM systems. Due to this limitation, commodity DRAM modules in the DDRx family rely on data pre-fetching, burst mode access, and double-data-rate optimizations to achieve high speed data transfers. The resulting excessive data pre-fetching and use of Delay locked loops (DLLs) greatly increases the energy cost of these commodity DRAM systems.

Alternatively, a few recent works have proposed utilizing dense wavelength division multiplexed (DWDM) based photonic interfaces to achieve high bandwidth in DRAM architectures [18], [19]. A DWDM optical fiber interface can carry approximately 64 wavelengths, which creates 64 parallel data-transfer channels on a single fiber to enable high bandwidth transfers. Moreover, these fibers have a pitch similar to that of the electrical pins, thus also ensuring high bandwidth density transfers. In recent years, innovations in Si-photonics technology have enabled on-chip Si waveguides, ring modulators, and ring detectors to function at 20Gbps data rate [20]. Advancements in CMOS technology have also enabled a single lane SerDes to operate at a high 25Gbps data rate [21]. Thus, cumulatively, single wavelength channels within fiber links can today operate at 20Gbps data rate, which corresponds to a 160GBps bandwidth per DWDM fiber. This implies that 8 such fibers can serve a cacheline of 512 bits in just 0.05ns (one beat of 20Gbps data rate). In contrast, 8 electrical pins would need a burst length of 64 requiring burst time of 32ns to serve a cacheline at 1GHz frequency and double data rate. This result highlights how a photonic interface can operate at many times higher speed and higher bandwidth, thereby alleviating pin-bandwidth limitations of a conventional electrical interface. Moreover, a photonic interface does not require on-die termination (ODT) and DLLs, which also greatly reduces the energy cost of the system. However, the energy overhead of the non-data-dependent power related to the thermal trimming of microring resonators should be considered in any cost-benefit analysis. Many studies [18][19][22][24][29][30] show that in spite of this extra energy overhead, photonic interfaces yield energy benefits compared to conventional electrical interfaces.

III. *3D-Wiz* Architecture: Overview

In this section, we elaborate on our proposed 3D-DRAM architecture called *3D-Wiz*. The 3D data array organization, subbank floorplan and TSV-based 3D signal routing for *3D-Wiz* are discussed in [24] in more detail. Fig. 1 shows a single 3D-Wiz module of 8 Gb capacity (at 45nm) that consists of a stack of 4 dies, and is divided into 4 identical ranks. Each rank consists of 8 identical banks. The stack of 4 DRAM dies is further stacked on top of one logic die. All of the global control logic (except subarray-level control) for the DRAM is integrated on the logic die. The logic die also contains all the opto-electrical circuits required to support a bidirectional, 256-bit wide, dense wavelength division multiplexing (DWDM) photonic data bus, which acts as an interface between the processor side memory controller and the *3D-Wiz* module.

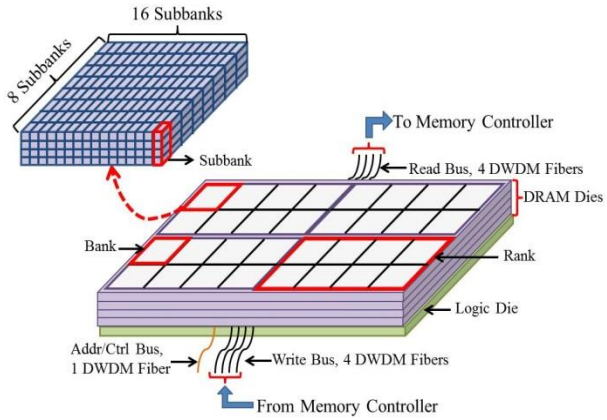

Fig. 1: Schematic of 3D-Wiz architecture and constituent elements.

A. *3D Organization of Data Array*

As discussed in [24], a bank of 256Mb size in *3D-Wiz* is divided into 128 smaller subbanks of size 2Mb each. A subbank is similar to a bank in functionality, except that it is smaller in size. Each subbank is folded across 4 die layers and consists of a total of 32 subarrays, 8 subarrays of which are on one layer. Fig. 2 shows the layout of a subbank. This array organization in *3D-Wiz* increases memory parallelism manifold, as each subbank can work independently. In commodity DRAMs, the *tFAW* power constraint significantly reduces bank-level parallelism in a rank [17]. As shown in Table 1, the precharge-activation energy of 3D-Wiz is about 10 times smaller than that of a commodity DDR3 module. Based on this result, we can conservatively estimate that the peak activation current drawn by *3D-Wiz* is about 8 times smaller than for commodity DDR3. Therefore, it is possible for a maximum of 8 subbanks to be concurrently activated in a 3D-Wiz module and have the same instantaneous current delivery capacity of as in DDR3. This translates into a tFAW (four-banks activation time window) value of zero in *3D-Wiz*. But, to put a limitation on the maximum number of activates in a rolling time window we consider the t32AW (32 activation window) metric. Accordingly, two groups of 8 activates each should be separated in time by at least tRRD=3ns, which translates into t32AW = 15ns.

B. *TSV-based 3D Signal Routing*

Fig. 2 shows the layout of a subbank and constituent subarrays, along with the TSV bus layout. All 32 constituent subarrays of a subbank are addressed in parallel, and they work in lockstep to serve a cache line. Each subarray has 4 data lines, which allows a subbank to serve a total of 128 data bits in one data burst. Therefore, *3D-Wiz* requires a burst length of 4 to serve a 64B cache line. The detailed floorplan of subbank is given in [24]. SuperContact [16] based TSVs are used in *3D-Wiz* for inter-layer vertical transfers. Compared to other 3D-DRAM architectures, *3D-Wiz* possesses performance and latency advantages due to its unique arrangement of subbanks and TSV bus described in detail in [24]. Briefly, in *3D-Wiz*, the decoder circuits for row and column address lines are located on the logic die. The decoded address lines are routed vertically via the TSV bus, which directly feeds the local word lines and column select lines of subarrays in a subbank. This arrangement eliminates the need for using on-die 2D routing, global word lines, global column select lines, and circuits connecting global lines with local lines, which alleviates the related latency and energy overhead. This arrangement yields an access time and row cycle time of 19.5ns and 25.1ns respectively for 3D-Wiz, which is significantly lower than in other 3D DRAM architectures, as discussed in more detail in subsection III.D.

As mentioned earlier, SuperContacts are responsible for feeding the address lines to all 32 parallel addressed subarrays of a subbank in *3D-Wiz*. Intuitively, this should require 32 sets of decoded row address and column address TSVs per subbank. Considering the pitch of 1.5 microns, the SuperContacts have an area footprint of 2.25μm^2. This would consume 9216μm^2 area for TSVs per subbank and 37.6mm^2 total TSV area per die. This is a very high area overhead and it seriously harms the overall area efficiency of the DRAM module. To overcome this challenge, we note that as all the subarrays in a subbank are addressed in parallel, we can use just one set of address TSVs. In this case, each address TSV, when it reaches a die layer, is used to feed in one 1:8 fan-out buffer after a repeater stage. Using one 1:8 fan-out buffer per address TSV on each die layer enables *3D-Wiz* to drive all 32 subarrays of a subbank using just one set of address TSVs. Given an area overhead of 5.2μm^2 for a fan-out buffer, the total area of all fan-out buffers is 6mm^2, which is about 6.3% of the total die area. The area of fan-out buffers and the area of one set of address TSVs per subbank sums up to be 10.6mm^2, which is about 30% of the area consumed by 32 sets of TSVs per subbank. Thus, this optimization in 3D-Wiz significantly improves area efficiency.

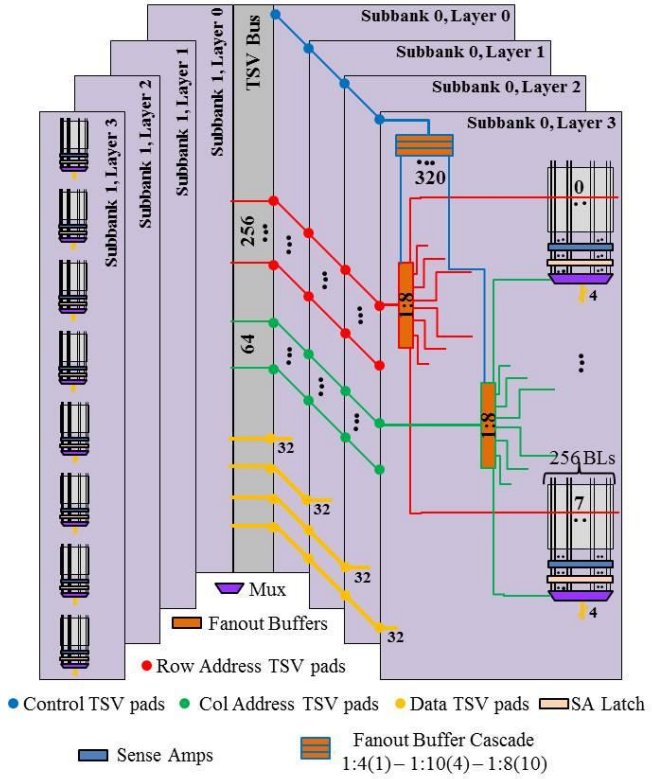

Fig. 2: Layout of subbank, TSV bus and constituent subarrays.

Moreover, as mentioned in [24], a TSV bus in *3D-Wiz* is shared between two subbanks, which requires one control signal to activate one of the two subbanks. This control signal needs to control all the fan-out buffers on each layer. As a subbank in *3D-Wiz* has a total of 320 fan-out buffers (Fig. 2) on one die layer, the control signal needs to be replicated 320 times. This can be achieved by using one control TSV per subbank (shown in blue color in Fig. 2) feeding in a cascade of fan-out buffers. In this cascade, one 1:4 fan-out buffer feeds in four 1:10 fan-out buffers which in turn feeds in ten 1:8 fan-out buffers, and at the end of the cascade we get 320 outputs each controlling one 1:8 fan-out buffer for address bits. The fan-out buffer cascade for the control signal occupies 15μm^2 area per die layer per subbank, which is not negligible compared to the

total area of 1:8 fan-out buffers for address bits. Also, the cascade yields 0.72ns delay and 35fJ energy consumption, which cumulatively realizes very low overhead control for subbanks in our *3D-Wiz* module.

C. High Bandwidth Photonic Interface

We now discuss the maximum theoretical bandwidth achievable with 3D-Wiz and justify the use of a high-speed and high-bandwidth photonic interface. As discussed earlier, the t32AW time is 15ns, which implies that about 32 subbanks per rank can be activated every 15ns. Thus, about 32 cachelines can be served per rank every 40ns (tRC+t32AW), which translates into a 51.2GBps bandwidth per rank and about a 205GBps bandwidth per 3D-Wiz module. A conventional electrical bus based interface of the DDRx family cannot support such a high bandwidth due to pin-bandwidth limitations. To address this issue, the Hybrid Memory Cube (HMC) from Micron proposes the use of high speed serial links at the interface [12]. This high speed link in HMC consists of several differential lanes as the fundamental building blocks. Each differential lane is claimed to achieve a maximum data transfer rate of 10Gbps. Thus, *3D-Wiz* would require about 164 such differential lanes to achieve a 200GBps data transfer rate across the interface. Unfortunately, this is a prohibitively large number of lanes and is impractical as it would cause serious packaging issues due to pin-limitations. Alternatively, the bandwidth requirement of *3D-Wiz* can be fulfilled by just 2 DWDM-based optical links, described in Section II. Therefore, we propose utilizing such a high bandwidth photonic interface for *3D-Wiz*.

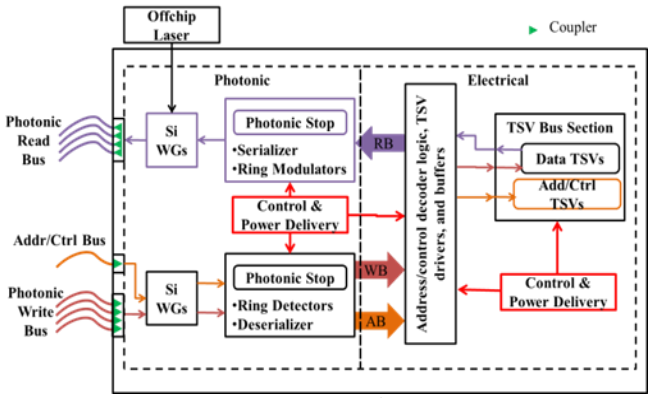

Fig. 3: Functional block diagram of the logic die in 3D-Wiz.

As shown in Fig. 3, the photonic interface in *3D-Wiz* is comprised of two links: one for reads and the other for writes. Each link consists of 4 unidirectional DWDM fibers. Each DWDM fiber in the read and write links supports 64 wavelengths, making the read and write buses 256-bits wide each. The photonic interface also uses one additional link consisting of a single DWDM fiber for transmitting address and control signals. The address/control fiber supports a total of 40 wavelengths, with 32 wavelengths for addresses and 8 wavelengths for control signaling. All the functionalities and circuits required to support this high bandwidth interface are accommodated on the logic die. Fig. 3 also shows how the logic die is functionally divided into two parts: photonic and electrical. The photonic part of the logic die does all of the optical to electrical (O-to-E) and electrical to optical (E-to-O) conversions; and optically interfaces with the memory controller on the processor chip. It consists of on-chip silicon waveguides (Si WGs), photonic couplers, and photonic stops. The electrical part of the logic die consists of address/control decoder logic, TSV drivers, buffers, and the control and power delivery network. It interfaces with the DRAM cell array through TSV buses.

D. Area, Latency and Energy Analysis

The area, timing and energy analysis for the *3D-Wiz* architecture was performed by enhancing the code for CACTI-3DD [23]. A similar analysis was conducted for other well-known 3D-DRAM architectures such as the 3D stacked photonic DRAM (3DSPDRAM) [19], 3D DRAM from Samsung (3DSAMS) [11], and the hybrid memory cube (HMC) from Micron [12]. The results of the analysis for these architectures were compared with the results for *3D-Wiz* and a commodity DDR3 DRAM architecture. The models of the aforementioned 3D-DRAM architectures were implemented in CACTI-3DD to the best of our knowledge using the technology parameters for the 45nm node. All TSVs in this study were modeled to be the SuperContacts from Tezzaron [16].

Table 1: Access time (tAC), row cycle time (tRC), activation energy (ActE), precharge energy (PreE), area efficiency (AE), and TSV area overhead per rank for different DRAM architectures

	tAC (ns)	tRC (ns)	ActE (nJ)	PreE (nJ)	AE (%)	TSV area overhead (mm^2)
3D-Wiz	19.5	25.1	0.78	0.62	42.5	3054×10^{-3}
HMC	22.5	38	1.27	1.13	54.5	44.7×10^{-3}
3DSPDRAM	26.3	40.3	1.08	1.01	42.6	0.8×10^{-3}
3D-SAMS	24.7	61.8	1.81	1.68	43.7	1.7×10^{-3}
DDR3	36.4	69.5	7.45	7.3	46.7	-

As can be seen from Table 1, 3D-Wiz demonstrates access latency of 19.5ns and row cycle time of 25ns, which are better than other architectures on average by about 27% and 52% respectively. The cost of a DRAM module depends on the area efficiency of the DRAM die. Therefore, area efficiency is a very important parameter to consider while designing a new DRAM architecture. Area efficiency of a DRAM die is defined as the percentage of the total die area which corresponds to the DRAM cell area. The total DRAM die area includes the area covered by peripherals, memory bus and TSVs in addition to the area covered by DRAM cells. A DRAM architecture with high area efficiency is thus able to store more useful information compared to a lower area efficiency DRAM architecture, within the same die area. It is evident from the results (Table 1) that despite the increase in the TSV area overhead for *3D-Wiz*, its area efficiency is comparable to most of the architectures listed in the table. This is because the TSV area overhead is counter-balanced by the area benefits obtained by optimized interconnects and relocation of decoder logic on the logic die in our architecture. However, the comparable area efficiency of *3D-Wiz* does not translate into great cost benefits in this case, as the cost of logic die has to be added in the total cost of the DRAM module. Nonetheless, the significant improvements in access time and energy consumption, as well as transfer bandwidth make *3D-Wiz* a promising architecture candidate for future 3D-DRAMs.

IV. Experimental Results

A. Experimental Setup

We performed trace-driven simulation analysis to compare *3D-Wiz* with other DRAM architectures. Memory access traces for the PARSEC benchmark suite [25] were extracted from detailed cycle-accurate simulations using gem5 [26]. We considered six different applications from the PARSEC suite: *Blackscholes*, *Dedup*, *Facesim*, *Ferret*, *Streamcluster*, and *x264*. We ran each PARSEC benchmark for a "warm-up" period of 100 million instructions and captured memory access traces from the subsequent 50 million instructions extracted. These memory traces were then provided as inputs to the DRAM simulator DRAMSim2 [27], which we modified heavily to model *3D-Wiz* and other DRAM architectures.

Table 2: Modeling parameters for the photonic interface [19]-[21]

Read laser power	1.22 mW
Write laser power	0.49 mW
Addr/Ctrl laser power	0.24 mW
Ring detector energy	44 fJ/bit
Thermal tuning power	50μW/ring
Ring modulator energy	47 fJ/bit
Modulator/detector delay	0.05 ns
SerDes delay	0.04 ns

Table 3: DRAMSim2 simulation configurations

3D-Wiz	8Gb module, 4 ranks, 1024 subbanks/rank, Stacked die count: 4 128-bit wide TSV data bus, Burst length 4
HMC	8Gb module, 16 banks/rank (vault), Stacked die count: 4 256-bit wide TSV data bus, Burst length 2
3DSPDRAM	8Gb module, 8 banks/rank, Stacked die count: 8 16-bit wide TSV data bus, Burst length 32, Single subarray access (SSA)
3DSAMS	8Gb module, 4 ranks, 8 banks/rank, Stacked die count: 4 32-bit wide TSV data bus, Burst length 16
DDR3	8Gb module, 1 rank, 8 banks/rank, 64-bit wide JEDEC data bus, Burst length 16

Table 2 gives the parameters used for modeling the photonic interface. We carefully modeled the differential lane based memory interface of HMC and the photonic interface of *3D-Wiz* to the best of our knowledge. Table 3 shows the memory configurations used in DRAMSim2 for the comparison across different DRAM architectures. A rank-based round-robin scheduling scheme and a closed page policy were used for all simulations. Performance, power consumption, and energy-delay product values for the memory subsystem were obtained from DRAMSim2. Performance was calculated as the inverse of average access latency. The results of the comparison study are discussed in the following subsection.

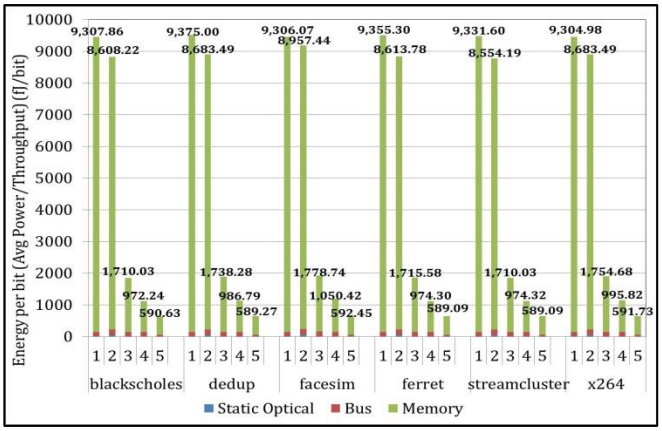

Fig. 4: Energy per bit values for PARSEC benchmarks. (1: DDR3, 2: 3DSPDRAM, 3: 3DSAMS, 4: HMC, 5: 3D-Wiz).

B. Results for PARSEC Benchmarks

This subsection presents the performance and energy bit values for all of the 3D DRAM designs shown in Table 5 obtained for simulation studies with PARSEC benchmark workloads. Fig. 4 shows energy per bit values for the various DRAM architectures across the PARSEC benchmarks. Energy per bit values are obtained by dividing the average power by throughput. The figure gives the total energy per bit which is a sum of static optical energy per bit, dynamic memory access energy (labeled as 'Memory') and energy consumed by the interface bus (labeled as 'Bus') that interfaces the DRAM module with the processor. It can be observed that 3D-Wiz consumes about 83.8% less energy per bit on average over all the other 3D-DRAM architectures. More specifically, 3D-Wiz consumes 93.2%, 92.8%, 65.8% and 43.6% less energy per bit on average over DDR3, 3DSPDRAM, 3DSAMS and HMC, respectively. The reason for the lower energy consumption in *3D-Wiz* is due to the smaller values of per access activation-precharge energy and relatively high throughput, the effect of which cumulates to minimize energy per bit.

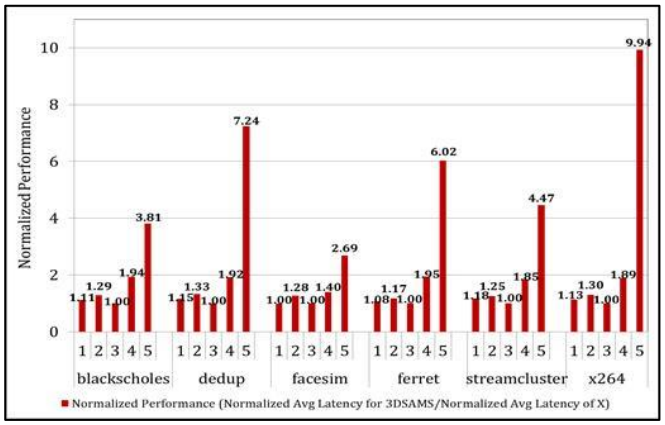

Fig. 5: Performance values for PARSEC benchmarks. (1: DDR3, 2: 3DSPDRAM, 3: 3DSAMS, 4: HMC and 5: 3D-Wiz).

Fig. 5 shows performance values for different DRAM designs normalized to the performance of 3DSAMS, across the PARSEC benchmarks. 3D-Wiz demonstrates about 4.4× greater performance on average over all the other 3D DRAM architectures. More specifically, 3D-Wiz demonstrates about 3.7×, 3.2×, 4.1×, and 2.3× greater performance values on average over DDR3, 3DSPDRAM, 3DSAMS and HMC respectively. As discussed in Section II and in [23], the fine-grained rank-level 3D partitioning of the data array in HMC better utilizes potential TSV bandwidth compared to the coarse-grained rank-level partitioning used in 3DSPDRAM and 3DSAMS. Due to this reason, HMC has an edge over 3DSPDRAM and 3DSAMS, which translates into a performance edge for HMC over 3DSPDRAM and 3DSAMS. The reason for the greater performance of *3D-Wiz* over other DRAM designs is the reduced RC loading of access paths in 3D-Wiz, which is a result of the smaller subarrays and elimination of global lines in its architecture.

In summary, our studies indicate that the utilization of 3D stacking technology and photonic interfaces can provide significant improvements in access time, energy consumption, as well as transfer bandwidth for DRAM modules. Our *3D-Wiz* architecture that was designed to intelligently integrate 3D stacking and high-bandwidth photonic interfaces provides one such very promising solution that may be able to overcome memory bottleneck concerns in future DRAM architectures.

V. CONCLUSIONS AND CHALLENGES

This paper introduced *3D-Wiz*, a novel high bandwidth and low latency 3D DRAM architecture. *3D-Wiz* uses TSVs and fan-out buffers to route address and data bits directly to individual subarrays and avoids global data lines. This in turn reduces random access latency and activation-precharge energy for the architecture. Also, *3D-Wiz* integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism than other 3D-DRAM architectures. Our experimental studies indicate that *3D-Wiz*

yields on average 83.8% and 4.4× improvements in energy per bit and performance, respectively, over the best known prior works in the area of 3D-DRAM architectures.

The significant improvements demonstrated by *3D-Wiz* position it as a promising architecture for future DRAMs. The performance of the *3D-Wiz* memory system can be further improved by using intelligent scheduling schemes and novel memory controller designs so that the greater parallelism of *3D-Wiz* architecture can be better exploited. However, there are some challenges that need to be overcome before the DRAM manufacturers can adopt mainstream mass production of 3D-stacked DRAMs with photonic interfaces. The TSV-based stacking of CMOS compatible silicon photonics die needs to be successfully and cost-effectively implemented. Also, the packaging technology of lateral and vertical photonic couplers needs to become more mature. Low cost methods for rapidly testing DRAM dies and CMOS compatible photonic dies before and after stacking need to be developed and demonstrated. Lastly, integration with waveguide [30] and free-space [31] based on-chip photonic networks would be an essential step towards reducing the memory and communication bottlenecks for future massively multi-core processor architectures.

ACKNOWLEDGMENTS

This research is supported by grants from Semiconductor Research Corporation (SRC), NSF (CCF-1252500, CCF-1302693), and AFOSR (FA9550-13-1-0110).

REFERENCES

[1] P. Kogge and K. Bergman, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," 2008.

[2] NEC, 64M-bit Virtual Channel SDRAM Data Sheet, NEC Corporation, 1998.

[3] Micron, DDR3 SDRAM MT41J2G4 Data Sheet, Micron Technology, Inc., 2011.

[4] S. Beamer et al., "Re-Architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics," in ISCA'10, June 2010.

[5] Y. H. Son, O. Seongil, R. Yuhwan, H. W. Lee and J. H. Ahn, "Reducing Memory Access Latency with Asymmetric DRAM Bank Organizations," in ISCA'13.

[6] O. Mutlu and T. Moscibroda, "Parallelism-Aware Batch Scheduling: Enhancing both Performance and Fairness of Shared DRAM Systems," in ISCA, 2008.

[7] E. Ipek, O. Mutlu, J. F. Martinez and R. Caruana, "Self-Optimizing Memory Controllers: A Reinforcement Learning Approach," in ISCA, 2008.

[8] J. Liu, B. Jaiyen, R. Veras and O. Mutlu, "RAIDR: Retention-Aware Intelligent DRAM Refresh," in ISCA, 2012.

[9] J. Mukundan, H. Hunter, K.-h. Kim, J. Stuecheli and J. F. Martinez, "Understanding and Mitigating Refresh Overheads in High-Density DDR4 DRAM Systems," in ISCA'13, 2013.

[10] D. H. Woo, N. H. Seong, D. L. Lewia and H. S. Lee, "An Optimized 3D-Stacked DRAM Architecture by Exploiting Excessive, High-Density TSV Bandwidth," in HPCA'16, Jan 2010.

[11] U. Kang, H.-J. Chung, H. Seongmoo, S.-H. Ahn, H. Lee, S.-H. Cha, J. Ahn, D. Kwon, J. H. Kim, J.-W. Lee, H.-S. Joo, W.-S. Kim, H.-K. Kim, E.-M. Lee, S.-R. Kim, K.-H. Ma, D.-H. Jang, N.-S. Kim, M.-S. Choi, S.-J. Oh, J.-B. Lee, T.-K. Jung, J.-H. Yoo and C. Kim, "8 Gb 3D DDR3 DRAM Using Through-Silicon-Via Technology," in JSSC, 2010.

[12] J. T. Pawlowski, "Hybrid Memory Cude (HMC)," in proceedings of Hot Chips 23, 2011.

[13] Y. H. Son, O. Seongil, Y. Ro, J. W. Lee and J. H. Ahn, "Reducing Memory Access Latency with Assymetric DRAM Bank Organization," in ISCA'13, 2013.

[14] C. Weis, I. Loi, L. Benini and N. Wehn, "Exploration and Optimization of 3D Integrated DRAM subsystems," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no. 4, 2013.

[15] B. Giridhar, M. Cieslak, D. Duggal, R. Dreslinski, H. M. Chen, R. Patti, B. Hold, C. Chakrabarti, T. Mudge and D. Blaauw, "Exploring DRAM Organizations for Energy-Efficient and Resilient Exascale Memories," SC, 2013.

[16] Tezzaron Semiconductor, [Online]. Available: http://www.tezzaron.com/about-us/our-technology-101/.

[17] B. Jacob, S. W. NG and D. T. Wang, Memory Systems: Cache, DRAM, Disk, Morgan Kaufmann, 2007.

[18] A. Hadke, T. Benavides, S. J. Ben Yoo, R. Amirtharajah and V. Akella, "OCDIMM: Scaling the DRAM Memory Wall Using WDM based Optical Interconnects," in HOTI, 2008.

[19] A. N. Udipi, N. Muralimanohar, R. Ralsubramonian and P. N. Jouppi, "Combining Memory and a controller with Photonics through 3D-Stacking to Enable Scalable and Energy-Efficient Systems," in ISCA'11, 2011.

[20] K. Bergman, L. P. Carloni, A. Biberman, J. Chan and G. Hendry, Photonic Network-On-Chip Design, Springer, May 2013.

[21] C. Zhong, C. Liu and F. Zhong, "25Gbps SerDes," IEEE HSSG, Orlando FL, 2007.

[22] S. Pasricha and N. Dutt, "Trends in Emerging On-Chip Interconnect Technologies," IPSJ Transactions on System LSI Design Methodology, vol. 1, 2008.

[23] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman and N. P. Jouppi, "CACTI-3DD: Architecture-level Modeling for 3D Die-stacked DRAM Main Memory," in DATE, 2012.

[24] I. Thakkar and S. Pasricha, "3D-Wiz: A Novel High Bandwidth, Optically Interfaced 3D DRAM Architecture with Reduced Random Access Time," in IEEE International Conference on Computer Design, Seoul, Korea, 2014.

[25] C. Bienia, S. Kumar, J. P. Singh and K. Li, "The PARSEC Benchmark Suit: Characterization and Architectural Implications," in PACT, Oct 2008.

[26] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill and D. A. Wood, "The gem5 Simulator," in Computer Architecture News, May 2011.

[27] P. Rosenfeld, E. Cooper-Balis and B. Jacob, "DRAMSim2: A Cycle Accurate Memory System Simulator," IEEE Computer Architecture Letters, 2011.

[28] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," in ISCA'35, 2008.

[29] S. Bahirat and S. Pasricha, "OPAL: A Multi-Layer Hybrid Photonic NoC for 3D ICs," in IEEE/ACM Asia & South Pacific Design Automation Conference (ASPDAC 2011), Yokohama, Japan, Jan 2011.

[30] S. Bahirat and S. Pasricha, "METEOR: Hybrid Photonic Ring-Mesh Network-on-Chip for Multicore Architectures," ACM Transactions on Embedded Computing Systems, vol. 13, no. 3s, March 2014.

[31] S. Bahirat, S. Pasricha, "HELIX: Design and Synthesis of Hybrid Nanophotonic Application-Specific Network-On-Chip Architectures", IEEE International Symposium on Quality Electronic Design (ISQED), Mar. 2014.